Classification Through Machine Learning Technique: C4.5 Algorithm based on Various Entropies

Seema Sharma
School of Information Technology,
UTD, RGPV, Bhopal ,M.P., India

Jitendra Agrawal
School of Information Technology,
UTD, RGPV, Bhopal ,M.P., India

Sanjeev Sharma
School of Information Technology,
UTD, RGPV, Bhopal ,M.P., India

ABSTRACT

Data mining is an interdisciplinary field of computer science and is referred to extracting or mining knowledge from large amounts of data. Classification is one of the data mining techniques that maps the data into the predefined classes and groups. It is used to predict group membership for data instances. There are many areas that adapt Data mining techniques such as medical, marketing, telecommunications, and stock, health care and so on. The C4.5 can be referred as the statistic Classifier. This algorithm uses gain radio for feature selection and to construct the decision tree. It handles both continuous and discrete features. C4.5 algorithm is widely used because of its quick classification and high precision. This paper proposed a C4.5 classifier based on the various entropies (Shannon Entropy, Havrda and Charvt entropy, Quadratic entropy) instance of Shannon entropy for classification. Experiment results show that the various entropy based approach is effective in achieving a high classification rate.

Keywords

Data Mining, Classification technique, Machine learning, Decision tree technique, C4.5 algorithm.

1. INTRODUCTION

In the recent year, huge amounts of data being collected and stored in databases everywhere across the globe mainly come from information industry and social sites. There are needed to extract and classify useful information and knowledge from large data. The data mining is most popular knowledge acquisition technique, which deals with this problem Data mining (DM) is used to extract the required data from large databases. It is the process of performing automated extraction and generating the predictive information from large database it is actually the process of finding the hidden information or patterns from the repositories [13]. Data mining consists of the various technical approaches including machine learning, statistic, and database system. The goal of the data mining process is to discover knowledge from large databases and transform into a human understandable format. The DM and knowledge discovery are essential component organization due to its decision making strategy.

Classification, regression and clustering are three approaches of data mining in which instances are grouped into identified classes [13]. Classification is a popular task in data mining especially in knowledge discovery and future plan, it provides the intelligent decision making, classification is not only study and examines the existing sample data but also predicts the future behaviour to that sample data. The classification includes two phases first is learning process phase in which analysis training data, the rule and pattern created. The second phase tests the data and archive the accuracy of classification patterns. [31] Clustering approach is based on unsupervised learning because there are no predefined classes. In this approach data may be grouped together as a cluster [2] [11]. Regression is used to map data item into a really valuable prediction variable.

Classification technique has various algorithms such as decision tree, nearest neighbour, genetic algorithm support vector machine (SVM) etc. [3]. In which Decision tree algorithm is widely used. In this paper, we examine the C4.5 decision tree algorithms based on various entropies and construct the classifier that classify the problem based on multiple entropies, involves the Shannon entropy, Quadratic entropy, havrda and Charvat entropy. In the rest of this paper gives machine learning technique Concepts in Section 2, related work in section 3, experimental Model in section 4. Performance metrics5, Comparative studies and result are presented in Section 6. The paper is concluded in Sections 7

2. MACHINE LEARNING TECHNIQUES

In the context of data mining, learning technique is generally classified as supervised and unsupervised learning technique both belong to machine learning technique. Classification is supervised learning that focus on the prediction based on known properties. A Classification task begins with a data set in which the class assignments are known. If the target or class label has numerical values then a predictive model uses. Regression algorithm is not a Classification algorithm. There are many classification algorithms some method are mostly used such as decision tree, Support vector machine, Naive Bayes, KNN etc.

2.1 Decision tree

The Decision tree is one of the classification techniques which is done by the splitting criteria. The decision tree is a flow chart like a tree structure that classifies instances by sorting them based on the feature (attribute) value. Each node in a decision tree represents a feature in an instance to be classified. All branches denote an outcome of the test, each leaf node hold the class label. The instances are classified from starting based on their feature value. Decision tree generates the rule for the classification of the data set. Three basic algorithms are widely used that are ID3, C4.5, and CART. [12] ID3 is an iterative Dichotomer 3. It is an older decision tree algorithm introduced by Quinlan Ross in 1986 [25]. The basic concept is to make a decision tree by using the top-down greedy approach. C4.5 is the decision tree algorithm generated by Quinlan [24]. It is an extension of ID3 algorithm.. C4.5 algorithm is widely used because of its quick classification and high precision. CART stands for Classification Regression Tree introduced by Bremen [4].The property of CART is that it is able to generate the regression tree. In Regression tree leaf node contains a real number instance of a Class. A decision tree classifier is built in two phases:

- A growth phase
- A prune phase

After the preliminary tree has been built that is 'growth phase', a sub-tree is created with the least estimated error rate, that is the 'prune phase'. The process of pruning the preliminary tree consists of removing small, deep nodes of the tree resulting from 'noise' contained in the training sample thus decreasing the risk of 'over fitting' and ensuring in a more precise classification of unknown data.

As the decision tree is being built, the goal at each node is to decide the split attribute (feature) and the split point that best divides the training instances belonging to that leaf. The value of a split point depends on how well it separates the classes. Numerous splitting indices have been proposed in the precedent to evaluate the quality of the split. The below fig 1 shows the decision tree of weather prediction data base.

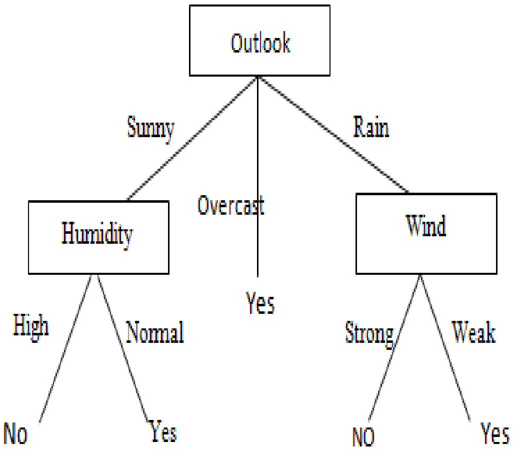

Fig.1 Decision Tree

2.2 K-Nearest Neighbor

The K-Nearest Neighbor (NN) is the simplest method of machine learning. It is a type of instance base learning in which object is classified based on the closest training example in the feature space. It implicitly computes the decision boundary however it is also possible to compute the decision explicitly. So the computational complexity of K NN is the function of the boundary complexity [9].The k-NN algorithm is sensitive to the local structure of the data set. The special case when k = 1 is called the nearest neighbor algorithm. The best choice of k depends upon the data set; larger values of k reduce the effect of noise on the classification [5] but make boundaries between classes less distinct. The various heuristic techniques are used to select the optimal value of K. KNN has some strong consistent results. As the infinity approaches to data, the algorithm is guaranteed to yield an error rate less than the Bayes error rate [5].

2.3 Support Vector Machine (SVM)

The support vector machine [SVM] is a training algorithm for classification rule from the data set which trains the classifier; it is then used to predict the class of the new sample. SVM is based on the concept of decision planes that define decision boundary and point to form the decision boundary between the classes called support vector threat as parameter. SVM is based on the machine learning algorithm, invented by vapnik in 1960's. and structure risk minimization principle to prevent over fitting. There are 2 key implementations of SVM technique: mathematical programming and kernel function [14]. It finds an optimal hyper plane between data point of different classes in a high dimensional space. We are concerned about two class classification, the classes being P and N for Yn= 1,-1, and by which we can extend to K class classification by using K two class classifiers. Support vector classifier (SVC) searches hyper plane but SVC is outlined so kernel functions are introduced in order to non line on decision surface.

2.3.1 SVC

Linear SVC used for data classification in which data is linearly separable. Let w is weight vector, b is base, Xn is the nearest data point. $w^T + b \geq 1$ for $x_n \in P$ .And $w^T x_n + b \leq -1$ for$x_n \in N$.

But A linear classifier not suitable for c class hypothesis. It can be used to learn nonlinear decision function space SVM can also be extended to learn non-linear decision functions [23]. The kernel function allows us to hyper plane without explicitly perform the calculation. [32]

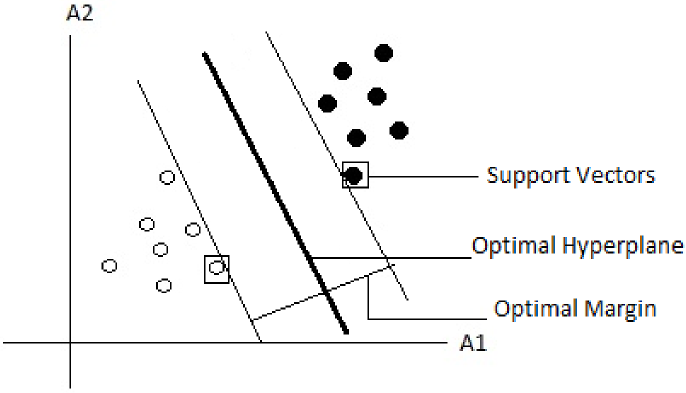

Fig2 Linear SVC

3. RELATED WORK

3.1 Decision tree Based Classification and Prediction

There are growing interest use of the decision tree learning algorithm to very large data set. So today globe various technology are proposed for improving decision tree. Sonal Agrawal [1] discussed classification decision tree approach in the field of education. Student data from a community college database has been taken and a variety of classification approaches have been performed and a comparative analysis has been done. The research work shows that the SVM is established as the best classifier with maximum accuracy and minimum root mean square error (RMSE). A Decision tree approach is proposed which may be taken as a significant basis of selection of student during several course programs.

HamidahJantan et al. [15] discussed an experimental study to discover the possible data mining classification technique for talent knowledge acquisition. Talent knowledge discovered from related databases can be used to classify the appropriate talent among employees. They use decision tree (C4. 5) Technique, to find the pattern of talent performance in the field of human resource. The author generates rules by evaluating using the unseen data in order to estimate the accuracy of the prediction result.

Phurivit et al [27] proposed real time intrusion detection approach using a supervised machine learning technique. They applied various well known machine learning techniques such as decision tree, back propagations, naive Bayesian classification and RBF NN to evaluate the performance of IDS. The author showed experimental results of the decision tree technique can outperform of IDS approach.

S. Ravikumar et al [26] discussed the machine learning approach in the field of inspection of machine component. This problem includes image acquisition pre-processing, feature selection and classification the author used Naive Baye algorithm and decision tree algorithm. And results showed that the accuracy of C4.5 is better than the other technique.

BahaSen [28] they developed models to predict secondary education placement test results, and using sensitivity analysis on those prediction models they find out the most important predictors. They proposed a decision tree algorithm for analysis of the success factor behind placement test that may help understand potentially improve the achievement.

Juan de ona et al [7] discusses road accidents, analysts many road safety is to identify the main factors that give to crash severity. They applied decision tree approach in the road safety field. The author's analysis the various decision tree algorithms and extracting the decision rule for road safety.

Jung Min Kim [16] studies, figure out what relationship the elements of meteorological changes have with the incidence of the five aggressive crimes during data mining. An analysis was made by the C4.5 algorithm of decision tree to verify what crimes occur according to the elements of the climate change.

3.2 Different entropy based C4.5 Algorithm for classification

Jiang Su and Harry Zhang [30] have discussed the decision tree method and also proposed a fast decision tree. They build a tree based on the conditional dependence assumption. Author shows that the performance and accuracy of the new approach is better than the C4.5 and less complexity as compare to C4.5 decision tree

In 2008 Tomasz Maszczyll and WlodzislawDuch [19] modify the C4.5 algorithm based on Tasallis and renyi entropy. After comparative analysis author is found that the modified C4.5 algorithm is better than the Shannon entropy based C4.5 algorithm. At basic of the decision tree algorithm ID3 are very famous and easy to classify but if classifying attribute which have many values then this algorithm are not beneficial.

Christiane FerreriaLemos Lima et al [18] describes the comparative study of the use of Shannon, Renyi and Tsallis entropies for designing decision tree. The goal of that paper is to find more efficient alternative entropy for the intrusion Tolerant system .The author show, the resultant tsallis and renyi entropy can be used to construct more compact and efficient decision tree.

In 2011 Mosonyi et al [21] proposed the Quantum Renyi Relative formula and relative capacity formula. The Shannon entropy is sensitive to noise sample and doesn’t work well in real work applications. So introduce the other measures of feature quality called the Rank mutual information.

Qinghua Hu et al [13] proposed the rank entropy based decision tree for monotonic classification. They apply rank mutual information which combined with the Shannon entropy. Author shows that if the training sample is consisting monotonically then performance is still good with presence of noisy data.

4. EXPERIMENTAL SETUP

Decision tree uses as a predictive model in which it maps the observations about an item to conclusions about the item's target value.C4.5 is one of the decision tree algorithms generated by Quinlan [24]. It is an extension of ID3 algorithm. C4.5 uses the Shannon Entropy.

Here, Shannon Entropy has been used to find information gain in C4.5 algorithm to calculate the Information Gain ratio contained by the data, which helps to make Decision Tree and to predict the targets. However, the results obtained from Shannon Entropy, are rather complex. Therefore, to minimize these problems we use other entropy such as R´enyi entropy, quadratic entropy, Havrda and Charvt entropy, and Taneja entropy instead of Shannon Entropy.

The architecture of the experimental model classification method has been divided into three phases
First Phase: Pre-Processing of data.

Second Phase In the second phase we apply various entropies to find the information gain ratio.
a) Shannon Entropy
b) Havrda and Charvt entropy
c) Quadratic entropy
d) R´enyi entropy
e)Taneja entropy
C4.5 algorithm is built in this phase based on the above entropies.

Third Phase: Output.

The classification includes a 3 step process

1) Model construction (Learning)
2) Model Evaluation (Accuracy)
3) Model Use (Classification)

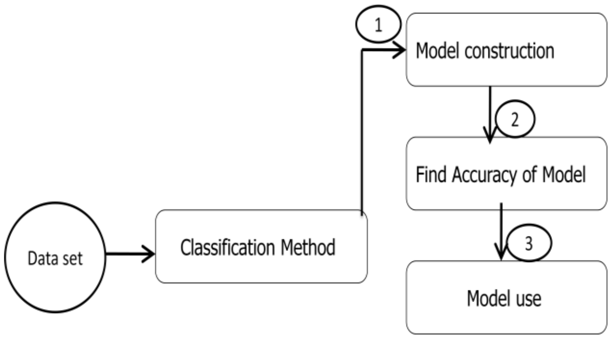

Fig 3 Classification Steps

All three steps are illustrated in fig 3. The first step is also known as learning. In this step learning is done on training sample data and construction of the model bases of learning is done. In the second step, the accuracy of model bases on test data is found. This phase treats as the testing phase. After testing phase, the accuracy of each model, used to classify the data set, is found. All these steps are essential for classification.

4.1 C4.5 algorithm

C4.5 is based on the information gain ratio that is evaluated by entropy. The information gain ratio measure is used to select the test features at each node in the tree. Such a measure is referred to as a feature (attribute) selection measure. The attribute with the highest information gain ratio is chosen as the test feature for the current node. Let D be a set consisting of (D_1… D_j) data instances. Suppose the class label attribute has m distinct values defining m distinct classes, Ci(for I = 1,…,m). Let Djbe the number of samples of D in class Ci. The expected information needed to classify a given sample is given by

$$\text{Splitinfo}_A(D)=- \Sigma \ (|Dj|/|D|)*\log((|Dj|/|D|)) \qquad (4.1)$$

$$\text{Gain ratio}(A)=\text{Gain}(A)\backslash\text{splitinfo}_A(D) \qquad (4.2)$$

Where

$$\text{Gain}= \text{Info}(D) - \text{InfoA}(D) \qquad (4.4)$$

$$\text{Info}(D)=- \Sigma \ \text{Pi} \log2(\text{Pi}) \quad \text{and} \qquad (4.3)$$

$$\text{Info}_A(D)=- \Sigma \ (|Dj|/|D|)* \text{Info}(Dj)$$

Where pi = probability of distinct class Ci,D =data Set, A=Sub attribute from attribute, (|Dj|/|D|)=act as weight of jthpartition. In other words, Gain (A) is the expected reduction in entropy caused by knowing the value of feature A.

Algorithm for Experimental Model

Input: dataset.
Output: classified output.

1. Take a data set as input.
2. If that set has more features then apply the feature selection technique (PCA) as pre-processing technique
3. Apply parallelism from step 4 to step 6.
4. Evaluate the entropy value and information gain ratio of all three entropies (Shannon, havrda and Charvat's entropy and quadratic entropy).
5. Construct the models separately using c4.5 algorithm based on various entropies.
6. Find the accuracy and execution time of each model and store the value in array.
7. Find a model that has maximum Accuracy.
8. If two have maximum accuracy then
9. Find a minimum execution time of the model that has maximum accuracy.
10. Classify by that model which has minimum execution time.
11. Else classification done by the model which has maximum accuracy.
12. End

4.2 Various Entropies

Entropy is a method of measuring randomness or uncertainty in a given set of data. For calculating the entropy of the file, our data set is a sequence of bytes in the file [22]. In the C4.5 algorithm information gain is derived from entropy. There are various types of entropies used in C4. 5 classification algorithms which are follow:

1. The Shannon entropy H (X) of a random variable X with a discrete probability distribution
P (i) = p1, p2, p3, p4…….. pk is given by:

$$H(x) = -\sum_{i=1}^{N} p(i)\log_2 p(i) \qquad (4.5)$$

2. 'Quadratic entropy' was initiated by I. Vajda. This entropy was first used in theoretical physics by Fermi. Consider the finite discrete random variable X with a complete probability collection.

$$P_i \geq 0 \ (i = 1 \ldots\ldots n), \quad \sum_{i=1}^{n} P_i = 1. \qquad (4.6)$$

Quadratic entropy is defined by the expression

$$H(X) = \sum_{i=1}^{n} P_i(1 - P_i). \qquad (4.7)$$

3. Havrda and Charvat's entropy [13] brings a further measure of entropy of the finite discrete random variable X with a complete probability mass function. The

Havrda and Charvat entropy of order α (α >0; α ≠ 1) is defined by the expression

4. Taneja entropy was introduced by I.J. Taneja in 1975 [29].A new formula for order c was proposed.

$$H(X) = \frac{\sum_{i=1}^{n} P_i^{\alpha} - 1}{2^{1-\alpha} - 1} \qquad (4.8)$$

$$H(x) = -\sum_{i=1}^{N} p(i)^{\alpha} \log 2p(i) \qquad (4.9)$$

5. R´enyi entropy [6] is parameterized according to the value of α, where α> 0 and α≠ 1, having Shannon entropy as the limit case. R´enyi entropy is given by:

$$R_\alpha(X) = \frac{1}{1-\alpha}\log\sum_{i=1}^{n} p_i^{\alpha} \qquad (4.10)$$

Where $\sum_{i=1}^{n} p\ (i)$ And $\lim_{\alpha\to 1} R_\alpha(X) = H(X)$.
α is a constant In this paper take α=0.25.

5. PERFORMANCE METRICS

5.1 Mean Squared Error

The mean squared error is probably the most important criterion used to evaluate the performance of a predictor. MSE measures the average of the squares of the errors. If y'is a vector of n predictions, and y is the vector of the true values, then the MSE of the predictor is:

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}(Y'_i - Y_i)^2 \qquad (5\ .1)$$

5.2 Cross validation

Cross-validation, also known as rotation estimation model [8][10][17]. This is a validation technique for measuring how the results of a statistical analysis will generalize to an independent data set. The fundamental approach, called k-fold CV, the training set is dividing into k smaller sets. The following is process is followed for each of the k "folds":

- A model is trained using $k-1$ of the folds as training data;
- The resulting model is validated on the remaining part of the data

In this research work uses 5-fold cross validation for testing.

5.3 Ttest2

Ttest2 performs a t-test of the null hypothesis of x, y data that are independent random samples from normal distributions with equal means and equal but unidentified variances against the alternative that the means are not equal. The end result of the test is returned in h. x and y need not be vectors of the same length and y can also be N-dimensional arrays or matrices. N-dimensional arrays x and y must have the same. The test treats Nan values as missing data, and ignores them.

$$[h, p, ci, stats] = ttest2\ (x, y) \qquad (5.2)$$

- h = 1 signifies a rejection of the null hypothesis at the 5% significance level. h = 0 signifies a failure to reject the null hypothesis at the 5% significance level
- The p value is the probability, under the null hypothesis, of observing a value as extreme or more extreme of the test statistic
- ci is 100* (1– alpha) % confidence interval on the difference of population means
- stats have the following fields:
  tstat — Value of the test statistic , df — Degrees of freedom of the test , sd — Pooled sample standard deviation or a vector with the sample standard deviations

6. EXPERIMENTAL RESULT

The main objective of this paper is to classify the dataset. We have eight standard datasets. As shown in Table 1, Bases on this data set valuate the accuracy, MSE of C4.5 algorithm based on different entropies shown in table 2,3 respectively. In this paper also comparative analysis among three entropies (Shannon, quadratic and Havrda & Charvt) based C4.5, C4.5 algorithm, KNN and SVM. Result has in table 4, fig 3 shows the graphically representation of output. And also apply the ttest2 on all method on six data sets.

6.1 Test Dataset

For evaluating the all eight C4.5 method based on various entropy; eight real world dataset were considered

Table 1. Properties Of Data Sets

Data set	Type of use	Origin	No. of Instance	No. of Feature	No. of Classes
Iris plants	Classification	Real world	150	4	3
Breast Cancer	Classification	Real world	277	9	2
Blood Transfusion Service Center Data Set	Classification	Business	748	5	2
Wine Quality	Classification , regression	Real world	1599 (red) 4898(wine	11	11
Yeast	Classification	Real world	1484	8	10
Pima Indians Diabetes	Classification	Real world	768	8	2
Thyroid Disease	Classification	Real world	7200	21	3
Ozone Level Detection	Classification	Real world	2536	73	2

6.2 Accuracy of Different Methods

There are five classification method testes by 5 fold cross validation on the eight data set table 2 Show the accuracy of each method on each data set and graphically shows by fig4.

Data set	C4.5 based on Shannon Entropy	C4.5 based on Havrda & Charvt entropy	C4.5 based on Quadratic entropy	C4.5 based on Renyi entropy	C4.5 Based on Taneja entropy
Iris	89.33%	92.67%	93.33%	90.00%	89.33%
Breast cancer	63.97%	93.09%	90.73%	63.82%	37.88%
Blood transfer	78.93%	75.44%	78.79%	76.24%	73.83%
Wine quality	54.67%	55.67%	55.68%	33.60%	44.26%
Yeast	33.58%	36.76%	29.53%	22.90%	8.24%
Pima	72.42%	66.93%	70.33%	65.23%	66.40%
Thyroid	97.35%	99.08%	99.10%	93.37%	92.58%
ozon layer	97.12%	97.04%	97.12%	97.12%	97.12%
Mean	69.90%	73.72%	73.05%	63.58%	64.56%

Table 2. Accuracy of Different Methods

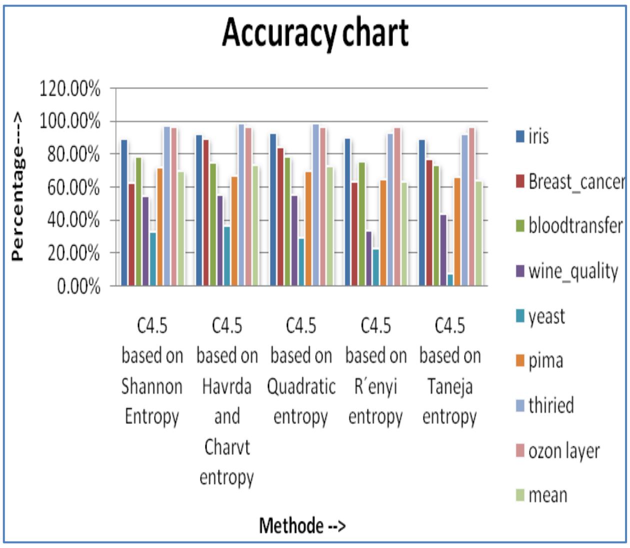

Fig.4 Accuracy Chart of Different Entropies Based C4.5 Algorithm

The below fig 4 show 5 different entropy based C4.5 methods for classification on eight different data sets that are , breast cancer , blood transfer , wine Quality yeast ,pima ,Thyroid,, ozon layer. This graphical representation show accuracy of each method on all data set. This graph show that Havrda and Charvat based C4.5 algorithm has maximum accuracy among all method that described above

6.3 Mean Squared Error of Different Methods

There are five classification method testes by 5 fold cross validation on the eight data set. In Section evaluates the Mean Squared Error of Different Methods, table 3 shows the MSE of each method on each data. Another work in this thesis is to built 3 entropies Shannon, Havrda and Charvt, Quadratic) Based C4.5 that apply all three entropy in parallel for building model and take best of one for Classify data. This technique compares with other machine learning techniques such as C4.5- algorithm, SVM (support vector machine), KNN (K-Nearest Neighbor) etc. Table 4 shows the accuracy of all machines learning technique.

Table 3. Mean Squared Error of Different Method

Data Set	Shannon Entropy based C4.5	Havrda and Charvt entropy based C4.5	Quadratic entropy based C4.5	R´enyi entropy based C4.5	Taneja entropy based C4.5
Iris	0.0033	0.0022	0.0017	0.0036	0.0033
Breast cancer	0.0357	0.0023	0.0105	0.0277	0.0154
Blood transfer	0.0109	0.0153	0.011	0.0148	0.0183
Wine quality	0.0414	0.0396	0.0394	0.0885	0.0639
Yeast	0.0888	0.081	0.0997	0.1204	0.1714
Pima	0.0157	0.0225	0.0184	0.0248	0.0233
Thiried	0.0001	0	0	0.0009	0.0011
ozon layer	0.00024	0.00025	0.00024	0.00024	0.00024
Mean	0.02451	0.02039	0.02261	0.03511	0.03711

6.4 Testing

Ttest2 performs on above machine learning algorithm. These tests carry out a t-test of the null hypothesis of x, y data that are independent random samples. The ttest carry out by equation (6.1) that is

$$[h, p, ci, stats] = ttest2(x, y) \qquad (6.1)$$

The result of ttest 2 is shown in table 5 .that consist degree of freedom, standard derivation value, confidential interval probability

Table 4 Accuracy of Machine Learning Algorithms

Data set	C4.5	C4.5 based on various entropy	SVM	KNN
Iris	89.33%	93.33%	92.00%	90.67%
Breast cancer	63.97%	93.09%	94.85%	57.50%
Wine quality	54.67%	55.68%	19.69%	42.57%
Yeast	33.58%	36.76%	50.14%	50.47%
Pima	72.42%	72.42%	73.46%	70.46%
thiried	97.35%	99.10%	94.28%	94.03%
Mean	**68.55%**	**75.06%**	**70.74%**	**67.62%**

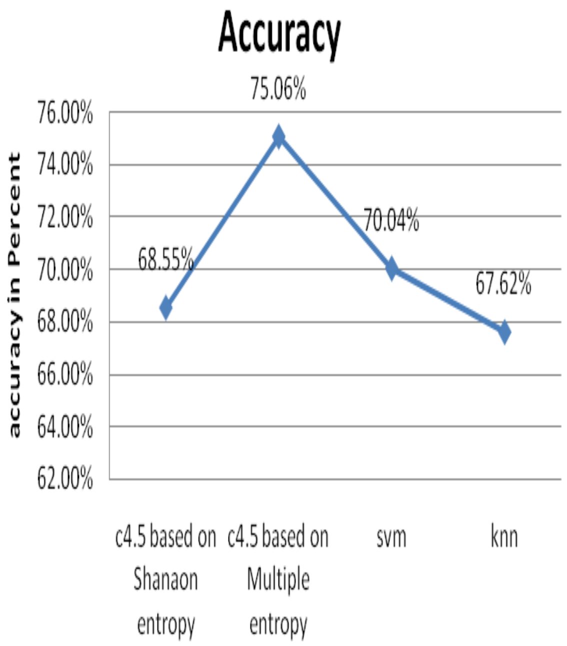

Fig.5 Mean Accuracy of Machine Learning Algorithm

Table 5. Ttest-2 of Machine Learning Techniques

Sample Data		Probability (p)	Confidence interval (CI)	Test statistic	Degree of freedom	Standard deviation
X	Y					
U	V	0.6497	-0.2447, 0.3749	0.4682	10	0.2408
U	W	0.7928	-0.3140, 0.4005	0.2699	10	0.2777
U	Z	0.5893	-0.223, 0.3719	0.5578	10	0.2312
V	W	0.8917	-0.3703, .3266	-0.1396	10	0.2709
V	Z	0.943	-0.2775, 0.2963	0.0727	10	0.223
W	Z	0.841	-0.3064, 0.3688	0.2059	10	0.2624

7. CONCLUSION

This experiment is performed over eight real datasets using the five methods namely C4.5 decision tree algorithm based on Shannon Entropy, C4.5 decision tree algorithm based on Havrda and Charvt entropy, C4.5 decision tree algorithm based on Quadratic entropy, C4.5 decision tree algorithm based on R´enyi entropy and C4.5 decision tree algorithm based on Taneja entropy. As shown in table 5, accuracy of Experimental Method based on three entropies is better than C4.5 algorithm. This paper also shows that comparative analysis between machine learning shown in above table.

Entropy Computation is used to create compact decision trees with successful classification. The size of the decision tree, the performance of the classifier is based on the entropy calculation. So the most precise entropy can be applied to the particular classification problem. The different entropies based approach can be applied in any classification problem. Such as detecting faults in industrial application, Medical diagnosis, loan approval, pattern recognition, classifying market trends etc. This thesis is a comparative study based on Shannon, R´enyi, quadratic, Havrda and Charvt, Taneja entropy and it also builds a model that takes Shannon, quadratic, and Havrda and Charvt entropy in parallel and produce more precise classification for data set and a result of this classification is comparable with the other machining learning techniques. This entropy based approach can be applied in real world classification problems.

8. REFERENCES

[1] Agarwal, S., Pandey, G. N., & Tiwari, M. D. Data Mining in Education: Data Classification and Decision Tree Approach.

[2] Merceron, A., & Yacef, K. (2005, May). Educational Data Mining: a Case Study. In *AIED* (pp. 467-474).

[3] Bakar, A. A., Othman, Z. A., & Shuib, N. L. M. (2009, October). Building a new taxonomy for data discretization techniques. In *Data Mining and Optimization, 2009. DMO'09. 2nd Conference on* (pp. 132-140). IEEE.

[4] Burrows, W. R., Benjamin, M., Beauchamp, S., Lord, E. R., McCollor, D., & Thomson, B. (1995). CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada. *Journal of applied meteorology*, *34*(8), 1848-1862.

[5] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, *13*(1), 21-27.

[6] Dasarathy, B. V. (1980). Nosing around the neighborhood: A new system structure and classification rule for recognition in partially exposed environments. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (1), 67-71.

[7] de Oña, J., López, G., & Abellán, J. (2012). Extracting decision rules from police accident reports through decision trees. *Accident Analysis & Prevention*.

[8] Devijver, P. A., & Kittler, J. (1982). *Pattern recognition: A statistical approach*(p. 448). Englewood Cliffs, NJ: Prentice/Hall International.

[9] Everitt,B. S., Landau, S., Leese, M.,& Stahl, D.Miscellaneous

Clustering Methods. *Cluster Analysis, 5th Edition*, 215-255.

[10] Geisser, S. (1993). *Predictive interference: an introduction* (Vol. 55).CRC Press.

[11] Han, J., Kamber, M., & Pei, J. (2006). *Data mining: concepts and techniques*. Morgan kaufmann.

[12] Horton, P., & Nakai, K. (1996, June). A probabilistic classification system for predicting the cellular localization sites of proteins. In *Ismb* (Vol. 4, pp. 109-115).

[13] Havrda, J., & Charvát, F. (1967). Quantification method of classification processes. Concept of structural $ a $-entropy. *Kybernetika*, *3*(1), 30-35.

[14] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Support Vector Machines. In *An Introduction to Statistical Learning* (pp. 337-372). Springer New York.

[15] Jantan, H., Hamdan, A. R., & Othman, Z. A. (2011, June). Talent knowledge acquisition using data mining classification techniques. In *Data Mining and Optimization (DMO), 2011 3rd Conference on* (pp. 32-37). IEEE.

[16] Kim, J. M., Ahn, H. K., & Lee, D. H. (2013). A Study on the Occurrence of Crimes Due to Climate Changes Using Decision Tree. In *IT Convergence and Security 2012* (pp. 1027-1036). Springer Netherlands.

[17] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI* (Vol. 14, No. 2, pp. 1137-1145).

[18] Lima, C. F. L., de Assis, F. M., & de Souza, C. P. (2010, May). Decision tree based on shannon, renyi and tsallis entropies for intrusion tolerant systems. In*Internet Monitoring and Protection (ICIMP), 2010 Fifth International Conference on* (pp. 117-122). IEEE.

[19] Maszczyk, T., & Duch, W. (2008). Comparison of Shannon, Renyi and Tsallis entropy used in decision trees. In *Artificial Intelligence and Soft Computing–ICAISC 2008* (pp. 643-651). Springer Berlin Heidelberg.

[20] Mathur, N., Kumar, S., Kumar, S., & Jindal, R. The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on Decision Tree.

[21] Mosonyi, M., & Hiai, F. (2011). On the quantum Renyi relative entropies and related capacity formulas. *Information Theory, IEEE Transactions on*, *57*(4), 2474-2487.

[22] Pareek, H., Eswari, P. R. L., Babu, N. S. C., & Bangalore, C. D. A. C. (2013). Entropy and n-gram Analysis of Malicious PDF Documents. *International Journal of Engineering*, *2*(2).

[23] Quinlan, J. R. (1987). Simplifying decision trees. *International journal of man-machine studies*,*27*(3), 221-234.

[24] Quinlan, J. R. (1993). *C4. 5: programs for machine learning* (Vol. 1). Morgan kaufmann.

[25] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

[26] Ravikumar, S., Ramachandran, K. I., & Sugumaran, V. (2011). Machine learning approach for automated visual inspection of machine components.*Expert Systems with Applications*, *38*(4), 3260-3266.

[27] Sangkatsanee, P., Wattanapongsakorn, N., & Charnsripinyo, C. (2011). Practical real-time intrusion detection using machine learning approaches.*Computer Communications*, *34*(18), 2227-2235.

[28] Şen, B., Uçar, E., & Delen, D. (2012). Predicting & analyzing secondary education placement-test scores: A data mining approach. *Expert Systems with Applications*, *39*(10), 9468-76.

[29] Sharma, B. D., & Taneja, I. J. (1975). Entropy of type (α, β) and other generalized measures in information theory. *Metrika*, *22*(1), 205-215.

[30] Su, J., & Zhang, H. (2006, July). A fast decision tree learning algorithm. In*Proceedings of the National Conference on Artificial Intelligence* (Vol. 21, No. 1, p. 500). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

[31] Jin, C., De-lin, L.,& Fen-xiang, M. (2009, July).An improved ID3 decision tree algorithm. In *Computer Science &education 2009. 4th International Conference on* (pp. 127-130). IEEE.

[32] Balagatabi, Z. N., & Balagatabi, H. N. (2013). Comparison of Decision Tree and SVM Methods in Classification of Researcher's Cognitive Styles in Academic Environment. *Indian Journal of Automation and Artificial Intelligence*, *1*(1),31-43